UDC 551.501.45:551.515.43:551.578.73(793)

Application of the Poisson and Negative Binomial Models to Thunderstorm and Hail Days Probabilities in Nevada

CLARENCE M. SAKAMOTO—*State Climatologist,*
National Weather Service, Reno, Nev.

ABSTRACT—Rare events such as thunderstorm or hail days often fit one of two distributions, the Poisson or the negative binomial. These two models were tested on monthly and annual thunderstorm days as well as on annual hail days at five locations in Nevada. A procedure for calculating efficient estimates of the parameters for the negative binomial distribution, utilizing the iterative process and the second-order polynomial model, is described. Results of the program applied to five sites in Nevada are discussed.

1. INTRODUCTION

The Poisson and the negative binomial distributions have been applied to rare events in meteorological and biological data (Bliss and Fisher 1953, Fisher 1941, Thom 1957, 1966). An excellent treatise on the history and properties of the negative binomial distribution is given by Williamson and Bretherton (1963). Generalized guidelines as to the adequacy of the two models have been discussed (Thom 1966), but, until pertinent tests are actually conducted, one cannot objectively determine which model is appropriate. Furthermore, calculations are laborious even with an electronic calculator. Iterative processes lend themselves to the use of the computers. Such is the case with the estimation of the parameters and probabilities for these two distributions.

The purpose of this study is to determine whether either the Poisson or the negative binomial model is adequate to describe the distribution of thunderstorm and hail days in Nevada. A computer program, developed to aid in the analysis, incorporated the following features:

1. Calculate the chi-square test of hypothesis to determine whether the Poisson or the negative binomial model is adequate.
2. Determine if the method of moments or the method of maximum likelihood should be used to estimate the parameters for the negative binomial distribution.
3. Calculate efficient estimates of the parameters by the maximum likelihood method without subjective graphical analysis.
4. Calculate probabilities for a selected number of thunderstorm or hail days without the use of the gamma function tables.

A thunderstorm day is defined as the occurrence day of at least one thunderstorm cloud (cumulonimbus) accompanied by lightning and thunder. It may or may not be accompanied by strong gusts of wind, rain, or hail. A hail day is a day when precipitation in the form of ice is produced by convective clouds. Small hail, usually a winter phenomenon, and hail have not been differentiated for the purpose of this study.

2. PROCEDURE

A test of hypothesis using the chi-square distribution with n-1 degrees of freedom (Thom 1966), is used to test whether the Poisson or the negative binomial distribution is adequate. It is given by

$$\chi^2_{n-1}=n\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i}-\sum_{i=1}^{n} x_i \tag{1}$$

where x is the number of event days, and n is the sample size or number of years. The test was made under the null hypothesis, H_0, that the occurrences or events follow the Poisson distribution with the same mean, that is, the sample frequency follows the population frequency. If the chi-square value with $n-1$ degrees of freedom at the 0.05 level of significance is exceeded, the null hypothesis is rejected and one should proceed with fitting the negative binomial distribution. A second-order curvilinear equation (Sakamoto 1972) simplifies the chi-square computation. (See app. A).

The probability function for the Poisson distribution is given by

$$f(x)=\frac{\mu^x e^{-\mu}}{x!}\qquad x=0,\,1,\,2\ldots,\,\infty \tag{2}$$

where μ is the population mean; x is the number of event days.

The negative binomial probability function is given by (Fisher 1941, Fisher 1953)

$$f(x)=\frac{(k+x-1)!}{x!\,(k-1)!}\left[\frac{p^x}{(1+p)^{k+x}}\right] \tag{3}$$

where x is the number of event days, and k and p are the parameters of the distribution. Equations (2) and (3) are evaluated following the procedure in appendix B. A

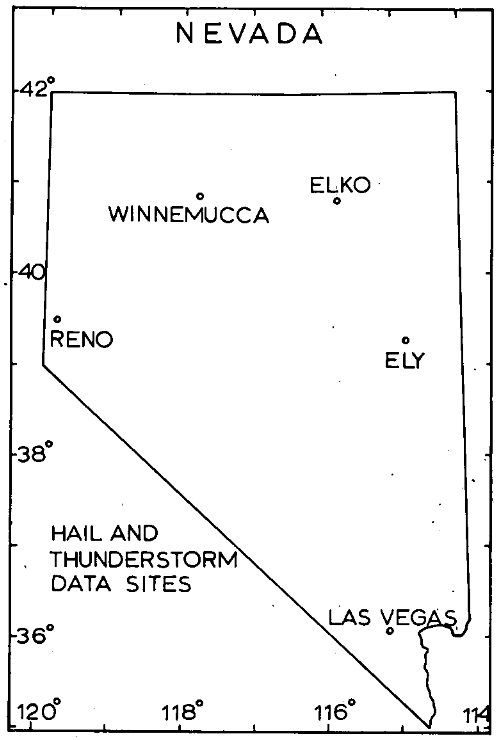

FIGURE 1.—Nevada data sites used in this study.

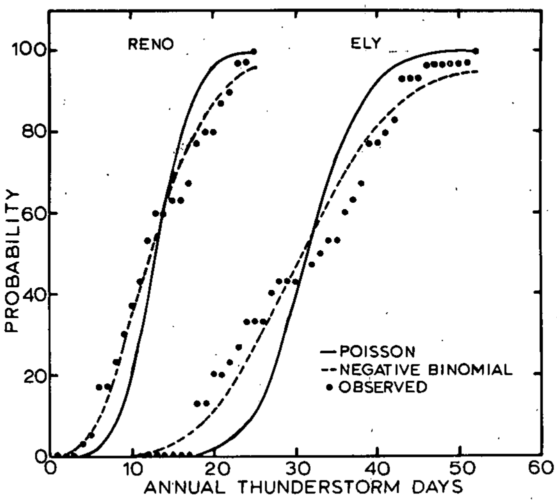

FIGURE 2.—Comparison of the observed cumulative frequencies of annual thunderstorm days with the theoretical probabilities for the Poisson and negative binomial distributions at Reno and Ely, Nev., 1941–70.

complete description of the final program and card format is given elsewhere (Sakamoto 1972). The program was subjected to monthly and annual thunderstorm days and annual hail days at Elko, Ely, Las Vegas, Reno, and Winnemucca, Nev. Outputs included the selected model, the probability, the selection of moments or maximum likelihood parameters estimate, the mean, the variance, and the sample size of each analysis.

3. DATA AND RESULTS

TABLE 1.—*Summary of model selection for thunderstorm and hail days in Nevada*

Period	Location				
	Ely	Reno	Elko	Winnemucca	Las Vegas
Jan.	*P**	None	*P*	*P*	*P*
Feb.	*N*	*P*	*P*	*P*	*P*
Mar.	*P*	*P*	*P*	*P*	*P*
Apr.	*P*	*P*	*P*	*P*	*N*
May	*N*	*N*	*N*	*N*	*P*
June	*N*	*N*	*P*	*N*	*P*
July	*N*	*N*	*P*	*N*	*N*
Aug.	*N*	*N*	*N*	*N*	*N*
Sept.	*N*	*P*	*N*	*N*	*N*
Oct.	*N*	*N*	*N*	*P*	*N*
Nov.	*P*	*N*	*P*	*N*	*P*
Dec.	*P*	*N*	*P*	*P*	*P*
Annual	*N*	*N*	*N*	*N*	*N*
Annual hail	*N*	*P*	*N*	*P*	*P*

*P=Poisson; N=negative binomial.

The *Local Climatological Data* (U.S. Department of Commerce 1942–71) and/or the station climatological record books for Elko, Ely, Las Vegas, Reno, and Winnemucca were used (fig. 1). Selection of the methods for monthly thunderstorm days and annual thunderstorm and hail days, based on the analysis of the chi-square test of hypothesis, are presented in table 1. In addition, observed versus calculated frequency curves for both the Poisson and the negative binomial model for sample locations are shown in figures 2 and 3. These figures show the annual thunderstorm days at Reno and Ely and the annual hail days at Elko and Ely, respectively. The data show the better fit of the negative binomial distribution in these cases.

Results in table 1 suggest that the model for estimating probabilities of selected number of thunderstorm days depends on the season and, hence, the climate of a particular region. For example, for the monthly data series (table 1), the negative binomial model is, with a few exceptions, adequate for the months May through October. The Poisson model is favored seasonally for the winter and early spring months at the majority of the sites.

The frequency of thunderstorm as well as hail days at both Elko and Ely are greater than at the other three locations. Geographically, Elko is located in the northeastern and Ely in the east-central portion of the State (fig. 1). Because the area is in the mean path of storm tracks, storms are more frequent.

Eleven cases in table 1 did not conform to the majority models. Seven of the 11 involved maximum probability differences of less than 0.023 between the two models. In the other four cases, the maximum difference between the two models was 0.108 involving, zero number of thunderstorm days.

For annual thunderstorm days, the negative binomial model was selected at all five sites. For annual hail days, however, only Ely and Elko, located in eastern Nevada, were associated with the negative binomial. As shown in

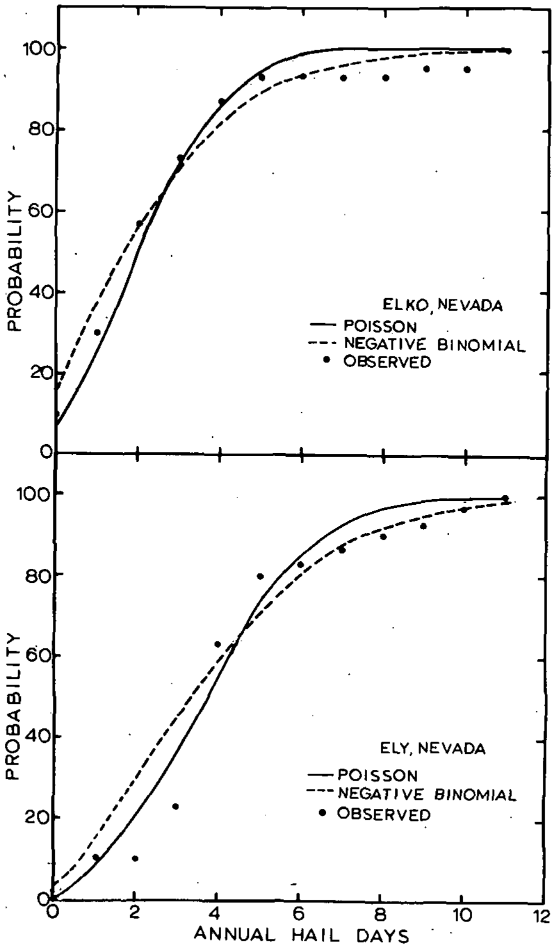

FIGURE 3.—Comparison of the observed cumulative frequencies of annual hail days with the theoretical probabilities for the Poisson and negative binomial distributions at Elko and Ely, Nev., 1941-70.

table 2, the hail day means at Ely and Elko are smaller than their variances. This is also true at all five sites for the annual thunderstorm days. Even though the mean at Winnemucca is smaller than its variance, the chi-square test of hypothesis showed that the Poisson model sufficed. This fact indicates that visually inspecting only the mean and the variance is insufficient to select the better model. In Nevada, the occurrence of thunderstorms in southern and western sections is much lower than in the northeastern area.

From the results in table 1 and the climatological association of these locations, it may be generally stated that, if the frequency is low, the Poisson is applicable, whereas, if the frequency is high, the negative binomial may be appropriate. However, as indicated above, objective tests are necessary to ascertain this general observation. Each station should be analyzed separately to determine the model that fits the data.

Calculated cumulative probabilities as well as observed frequencies for annual thunderstorm and annual hail days are shown in tables 3 and 4, respectively. The Kolmogorov-Smirnov test (Massey 1951) showed that the calculated probabilities fitted the observed data at the 0.10 level of significance, indicating that the selected models are satisfactory.

TABLE 2.—*Mean and variance of annual thunderstorm and annual hail days at five locations in Nevada (1941-70)*

Locations	Thunderstorm		Hail	
	Mean	Variance	Mean	Variance
Elko	24. 23	39. 47	2. 67	6. 09
Ely	31. 97	97. 69	4. 27	7. 24
Las Vegas	13. 47	25. 84	0. 13	0. 12
Reno	13. 50	37. 22	1. 17	1. 11
Winnemucca	15. 43	47. 08	2. 40	3. 14

One may interpret these tables as follows: The computed probabilities for 0 number of thunderstorm or hail days is the chance of none occurring at each site. For example, in table 4 at Las Vegas, the probability of no hail is 0.875. The probability of exactly x number of hail days can also be obtained. For example, $x=4$ at Elko is 0.815 minus 0.710 or 0.105. The probability of 4 or less days is 0.815 and the probability of more than 4 hail days is 1.000 minus 0.815 or 0.185.

4. CONCLUDING REMARKS

Although this study involved only thunderstorm and hail days, the models also have potential applications to other rare events such as tornadoes, hurricanes, biological data, etc. The procedure discussed above yields a satisfactory estimate of the parameters by the maximum likelihood method and eliminates the tedious process of estimating by eye the value of k at $L_2=0$.

In Nevada, the negative binomial model is adequate, in general, for monthly thunderstorm days from May through October, for annual thunderstorm days throughout Nevada, and for annual hail days in northeastern and east-central Nevada. On the other hand, the Poisson distribution is preferred for monthly thunderstorm days from November through April, as well as for annual hail days in southern and western Nevada. This indicates that climatic differences affect the utility of these two models and suggests that each climatically different site should be analyzed separately to determine the proper model that fits the data.

APPENDIX 1

In the computer program, a second-order curvilinear equation was developed to relate degrees of freedom and chi-squared values. Chi-squared values can be found in many elementary statistics texts. The established equation (Sakamoto 1972) between these two variables is

$$Y=4.54921+1.41672D-0.0036744D^2$$

where Y, the chi-squared value at the 0.05 level of significance, and D, the degrees of freedom, explained 0.999994 of the variation of the data about the curve.

APPENDIX 2

Expressed in natural logarithms, the Poisson density function is

$$\ln P=x \ln \bar{x}-\ln x!-\bar{x} \tag{4}$$

TABLE 3.—*Calculated (C) and observed (O) cumulative probabilities of annual thunderstorm days at five locations in Nevada (1941–70)*

No. days	Locations									
	Elko		Ely		Las Vegas		Reno		Winnemucca	
	C	O	C	O	C	O	C	O	C	O
0	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000
1	.000	.000	.000	.000	.001	.000	.003	.000	.002	.000
2	.000	.000	.000	.000	.002	.000	.009	.000	.007	.033
3	.000	.000	.000	.000	.007	.000	.021	.000	.017	.033
4	.000	.000	.000	.000	.018	.033	.042	.033	.032	.067
5	.001	.033	.000	.000	.037	.033	.072	.067	.055	.100
6	.002	.033	.000	.000	.066	.067	.112	.167	.086	.133
7	.005	.033	.000	.000	.108	.167	.162	.167	.123	.133
8	.010	.033	.001	.000	.162	.167	.219	.233	.167	.133
9	.018	.033	.002	.000	.227	.200	.283	.300	.217	.167
10	.030	.033	.003	.000	.300	.367	.349	.367	.272	.233
11	.048	.067	.005	.000	.380	.367	.418	.433	.329	.267
12	.072	.100	.009	.000	.460	.433	.485	.533	.387	.333
13	.103	.133	.013	.000	.540	.500	.551	.600	.446	.400
14	.143	.133	.020	.000	.615	.567	.612	.600	.503	.467
15	.189	.133	.028	.000	.684	.667	.669	.633	.558	.467
16	.242	.200	.039	.000	.745	.733	.720	.633	.610	.567
17	.300	.300	.053	.000	.798	.800	.765	.667	.658	.667
18	.362	.300	.070	.133	.842	.833	.805	.767	.702	.700
19	.426	.333	.090	.133	.879	.867	.839	.799	.743	.733
20	.491	.433	.113	.200	.908	.867	.869	.799	.779	.733
21	.554	.533	.139	.200	.931	.933	.894	.867	.811	.833
22	.615	.567	.168	.233	.949	.967	.914	.899	.840	.867
23	.672	.700	.200	.267	.963	.997	.932	.966	.865	.900
24	.724	.700	.235	.333	.973	1.000	.946	.966	.886	.900
25	.770	.800	.272	.333	.981		.957	1.000	.905	.900
30	.924	.900	.475	.433	.998		.988		.964	1.000
35	.981	1.000	.667	.533			.997		.987	
40	.996		.814	.767					.996	
45			.906	.933						
50			.942	.967						
55			.942	1.000						

TABLE 4.—*Calculated (C) and observed (O) cumulative probabilities of annual hail days at five locations in Nevada (1941–70)*

No. days	Locations									
	Elko		Ely		Las Vegas		Reno		Winnemucca	
	C	O	C	O	C	O	C	O	C	O
0	0.160	0.100	0.044	0.000	0.875	0.867	0.311	0.333	0.091	0.100
1	.370	.300	.147	.100	.992	1.000	.674	.633	.308	.367
2	.561	.567	.292	.100	1.000		.887	.867	.570	.600
3	.710	.733	.450	.233			.969	1.000	.779	.767
4	.815	.867	.596	.633			.993		.904	.867
5	.886	.933	.717	.800			.999		.964	.933
6	.931	.933	.809	.833					.988	.967
7	.959	.933	.876	.867					.997	1.000
8	.976	.933	.922	.900						
9	.986	.967	.952	.933						
10	.992	.967	.971	.967						
11	.995	1.000	.983	1.000						
12			.990							
13			.994							
14			.997							
15										

where P is the probability of exactly x event days, and $\bar{x}$ is the sample mean. Expressed in natural logarithms, the negative binomial density function is

$$\ln P = k \ln\left(\frac{1}{1+p}\right) + \ln K + x \ln \frac{p}{p+1} \qquad (5)$$

where P is the probability of x event days, k and p are the parameters of the distribution, and K is defined as

$$K = \frac{(k+x-1)!}{x!\,(k-1)!}. \qquad (6)$$

The parameters k and p are initially estimated by the method of moments (Fisher 1941, Thom 1957). That is,

$$k^* = \frac{\bar{x}^2}{s^2 - \bar{x}} \qquad (7)$$

and

$$p^* = \frac{s^2 - \bar{x}}{\bar{x}} \qquad (8)$$

where $\bar{x}$ and s^2 are the sample mean and variance, respectively. The asterisk indicates the moments estimator.

The moments estimator is not always efficient. This means that the estimator of the parameters k^* and p^* by the method of moments may not yield the maximum amount of information from the sample. In other words, the estimates may not provide minimum variance. To compensate for this loss of information, one needs a larger sample size if an increase in efficiency is desired. Since climatic records are often limited by the sample size, it is important to obtain an efficient estimate. In this study, if the variance of the moments estimator is more than 10 percent larger than the variance of another method, the efficiency is less than 90 percent and a more satisfactory estimate must be determined. Fisher (1941) has shown under what conditions a moments estimator is efficient and provides the following equation for testing the efficiency of the moments parameters:

$$C = \left(1 + \frac{1}{p^*}\right)(k+2). \qquad (9)$$

If $C<20$, the efficiency of the moments parameters is less than 90 percent and, therefore, the method of maximum likelihood should be used to estimate p and k; if $C>20$, the moments parameters are considered efficient.

Another method of estimation that provides the estimators for the negative binomial with minimum variance is the principle of maximum likelihood (Fisher 1941, Thom 1957). This procedure (e.g., Anderson and Bancroft 1952) maximizes the likelihood function and provides estimates that will maximize the probability of obtaining the function. Three steps are followed to determine the maximum likelihood estimator:

1. Let $f(x_i;\, k,\, p)$ be the distribution function. [See eq (3).]
2. Let $L = \ln\,[f\,(x_i;\, k,\, p)]$.
3. Maximize L with respect to p and k by partially differentiating the function with respect to p and k as follows:

$$\frac{\partial L}{\partial p} = 0; \qquad \frac{\partial L}{\partial k} = 0.$$

Table 5.—*Comparison of parameter* k *estimates by method of maximum likelihood (MXL), method of moments (MOM) and by eye for the negative binomial distribution*

Period	Elko			Ely			Las Vegas			Reno			Winnemucca		
	MXL	MOM	EYE	MXL	MOM	EYE	MXL	MOM	EYE	MXL	MOM	EYE	MXL	MOM	EYE
May	2. 228	1. 819	2. 226	4. 013	4. 067	4. 017				2. 277	3. 447	2. 273	1. 377	1. 573	1. 366
June				3. 499	3. 197	3. 497				1. 784	1. 739	1. 779	2. 042	2. 560	2. 040
July				3. 047	3. 927	3. 037		6. 750		3. 060	4. 522	3. 064	1. 855	2. 361	1. 849
Aug.	3. 315	4. 735	3. 316	5. 831	5. 474	5. 831	2. 180	2. 614	2. 174	1. 035	1. 109	1. 037	1. 227	1. 652	1. 222
Sept.	1. 833	2. 133	1. 833	3. 368	3. 333	3. 373	1. 704	2. 169	1. 700				1. 960	2. 138	1. 956
Oct.	0. 840	1. 065	0. 840	0. 902	1. 044	0. 896	0. 382	0. 271	0. 381	0. 259	0. 190	0. 247			
Annual		24. 233			15. 548			14. 652		7. 282	7. 682	7. 282	6. 236	7. 526	6. 241

Let L be equal to the likelihood function in the product rule form; that is,

$$L=\prod_{i=1} f(x_i;\, p,\, k). \qquad (10)$$

Substituting eq (3) in eq (10) and maximizing L with respect to p by partially differentiating the natural logarithm of L with respect to p, we have as the first equation (Fisher 1953, Thom 1957),

$$\bar{x}=kp.$$

Differentiating partially the natural logarithm of L with respect to k and, for convenience, letting it be equal to L_2, gives the second equation (Haldane 1941, Fisher 1953),

$$L_2=kn\ \ln\left(1+\frac{\bar{x}}{k}\right)-\left[(g_1+g_2+\ldots+g_x)\right.$$
$$\left.+\frac{k}{k+1}(g_2+g_3+\ldots+g_x)+\ldots+\frac{k}{k+x-1}(g_x)\right]=0 \qquad (12)$$

where g_1, g_2, . . ., g_x are the observed frequencies or counts of x number of thunderstorm or hail days from 1 to the highest number.

The value of k at $L_2=0$ is the final estimate of the maximum likelihood estimator of k. This is often determined iteratively by trial and error and graphically plotting the values to obtain an estimate of the final k. This process can become tedious. In this study, it is suggested that the value of k at $L_2=0$ can be solved rapidly by utilizing the second-order polynomial (curvilinear) equation by regressing values of L_2 on increment values of k. The curvilinear model is

$$L_2=C+Bk+Ak^2 \qquad (13)$$

where A, B, and C are constants.

The computer procedure for this method follows:

1. Find an initial estimate of k^* by the method of moments [eq (7)].

2. Increment and decrement iteratively k^* by a selected amount in eq (12) until values close to zero are obtained. In most cases, this means both positive and negative values of L_2 for iterative values of k. In this study, nine values of k with its corresponding values of L_2 were used in eq (13). If both positive and negative values of L_2 are not attained in the reiteration, the six values closest to zero are used.

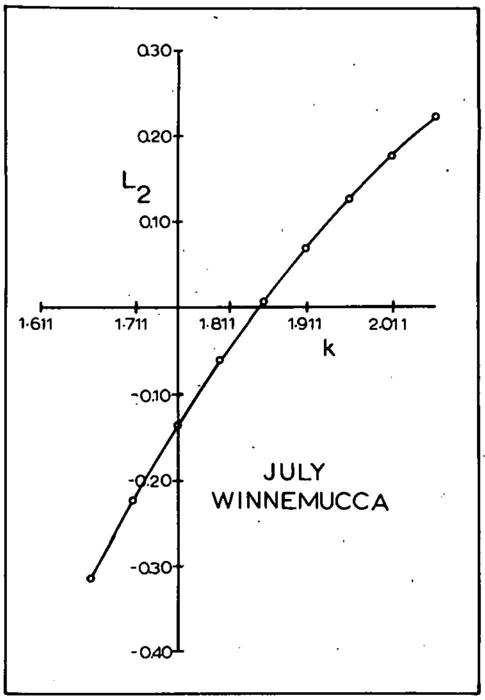

Figure 4.—Sample plot of computer-selected values of k and L_2 using the curvilinear model.

3. Solve eq (13) by least-squares procedure.

4. Set the derived curvilinear equation to zero and solve for k by the quadratic equation,

$$k=(-B\pm\sqrt{B^2-4AC})/2A.$$

Only positive values are determined.

5. Using eq (11), solve for p.

This procedure, involving the curvilinear model to calculate the efficient estimate of k, was attempted after repeated trials of curve-fitting the data. This procedure eliminates the tedious process of fitting the curve by eye. Figure 4 is an example of the plot between L_2 versus selected values of k (the abscissa) and the resultant curvilinear line for July thunderstorm days at Winnemucca.

Results of the procedure for estimating the parameter k when L_2 [eq (12)] is zero and that for estimating k by graphical (eye) procedure is shown in table 5. Estimates of the parameter by the moments method is also included. Excellent agreement is shown. It is concluded that, at least for the data analyzed, the procedure utilized in this

Nevada study is both a reliable and a rapid method for calculating the parameters of the negative binomial distribution by the maximum likelihood method.

ACKNOWLEDGMENTS

The author is grateful to his wife, Winifred, for assisting with this study and with the computer programming. Young Koh, statistician and computer consultant at the University of Nevada, aided our efforts in programming. Appreciation is expressed to Marvin Magnuson, National Weather Service, Western Region climatologist, NOAA, and to J. T. Lee, project leader with the National Severe Storms Laboratory, NOAA, for their constructive review and criticisms.

REFERENCES

Anderson, R. L., and Bancroft, T. A., *Statistical Theory in Research*, McGraw-Hill Book Company, Inc., New York, N.Y., **1952**, 339 pp.

Bliss, C. I., and Fisher, Ronald A., "Fitting the Negative Binomial Distribution to Biological Data," *Biometrics*, Vol. 9, No. 2, Raleigh, N.C., June **1953**, pp. 176–196.

Fisher, Ronald A., "The Negative Binomial Distribution," *Annals of Eugenics*, Cambridge University Press, London, England, Vol. 11, **1941**, pp. 182–187.

Fisher, Ronald A., "Note on the Efficient Fitting of the Negative Binomial," *Biometrics*, Vol. 9, No. 2, Raleigh, N.C., June **1953**, pp. 197–200.

Haldane, J. B. S., "The Fitting of Binomial Distributions," *Annals of Eugenics*, Cambridge University Press, London, England, Vol. 11, No. 13, **1941**, pp. 179–181.

Massey, Frank J. Jr., "The Kolmogorov-Smirnov Test for Goodness of Fit," *American Statistical Association Journal*, Vol. 46, No. 253, Mar. **1951**, pp. 68–78.

Sakamoto, Clarence M., "Thunderstorms and Hail Days Probabilities in Nevada," NOAA *Technical Memorandum* NWSTM WR–74, Apr. **1972**, 25 pp.

Thom, Herbert C. S., "The Frequency of Hail Occurrence," *Archiv für Meteorologie, Geophysik und Bioklimatologie*, Springer-Verlag, Inc., New York, N.Y., Series B, Band 8, 2 Heft, **1957**, pp. 185–194.

Thom, Herbert C. S., "Some Methods of Climatological Analysis," *Technical Note* No. 81, World Meteorological Organization, Geneva, Switzerland, **1966**, pp. 30–34.

U.S. Department of Commerce, NOAA, Environmental Data Service, *Local Climatological Data* for Reno, Ely, Elko, Las Vegas, and Winnemucca, Nevada, 1941-1970, Asheville, N.C., **1942–1971**.

Williamson, E., and Bretherton, M. H., *Tables of the Negative Binomial Probability Distribution*, John Wiley and Sons, New York, N.Y., **1963**, pp. 7–15.

[*Received March 27, 1972; revised March 2, 1973*]